Improving Few-Shot Learning with Multilingual Transfer and Monte Carlo Training Set Selection

Antonis Maronikolakis[△*] **Paul O'Grady**[▽] **Hinrich Schütze**[△] **Matti Lyra**[▽]
[△]Center for Information and Language Processing, LMU Munich
[△]Munich Center for Machine Learning
[▽]Zalando SE
antmarakis@cis.lmu.de

Abstract

In industry settings, machine learning is an attractive tool to automatize processes. Unfortunately, annotated and high-quality data is expensive to source. This problem is exacerbated in settings spanning multiple markets and languages. Thus, developing solutions for multilingual tasks with little available data is challenging. Few-shot learning is a compelling approach when building solutions in multilingual and low-resource settings, since the method not only requires just a few training examples to achieve high performance, but is also a technique agnostic to language. Even though the technique can be applied to multilingual settings, optimizing performance is an open question. In our work we show that leveraging higher-resource, task-specific language data can boost overall performance and we propose a method to select training examples per their average performance in a Monte Carlo simulation, resulting in a training set more conducive to learning. We demonstrate the effectiveness of our methods in fashion text reviews moderation, classifying reviews as related or unrelated to the given product. We show that our methodology boosts performance in multilingual (English, French, German) settings, increasing F1 score and significantly decreasing false positives.

1 Introduction

In real-life settings, machine learning methods are being applied to automate and improve processes, from content moderation to search query filtering. Large pretrained language models have the potential to bring further improvements, at the cost of data resources, with high quantities of labeled data required for effective training. Collecting, cleaning, annotating and analyzing data is an expensive, time-consuming and challenging task.

With recent advancements in modeling, it has been shown that large language models exhibit few- and zero-shot capabilities, able to solve tasks with little or no data (Wei et al., 2022). For example, with PET (Schick and Schütze, 2021a), models can be finetuned on a task using only a few training examples. Annotating a small number of examples is desirable and applicable in academic as well as real-life settings.

Model	**F1**	**FPR**
DistilDE	73.0%	30.7%
Set_{111}	73.4%	39.5%
Set_{MC}	74.7%	28.5%
mBERT_{all}	67.5%	16.2%
$\text{mBERT}_{all/MC}$	70.2%	25.9%

Table 1: Comparison (in the German setting) between performance of a production DistilBERT model (DistilDE) and the best-performing few-shot model (Set_{111}), as well as our models developed with our multilingual transfer learning (mBERT_{all}) and Monte Carlo (Set_{MC} and $\text{mBERT}_{all/MC}$) sampling methods. With our methods we can find a better balance between overall performance (F1 score) and the false positive rate.

An added complication in many scenarios is the need to develop solutions for multilingual settings. Annotating data, developing and evaluating models is increasingly more challenging when there are multiple languages to consider. This is exacerbated in settings where certain markets are larger and higher-resource than other markets. **Developing solutions that are scaleable to both high- and low-resource markets is challenging.**

We work in the domain of **customer text reviews on a fashion platform**. When a customer leaves a review on one of their purchased products, the review goes through moderation to verify whether it abides by the platform's code of conduct. For example, a review may be rejected because it contains offensive content or personal data. In our work, we are focusing on the task of identifying

[*] Work was performed while at Zalando SE.

Proceedings of the 2023 CLASP Conference on Learning with Small Data, pages 1–10
September 11–12, 2023. ©2023 Association for Computational Linguistics

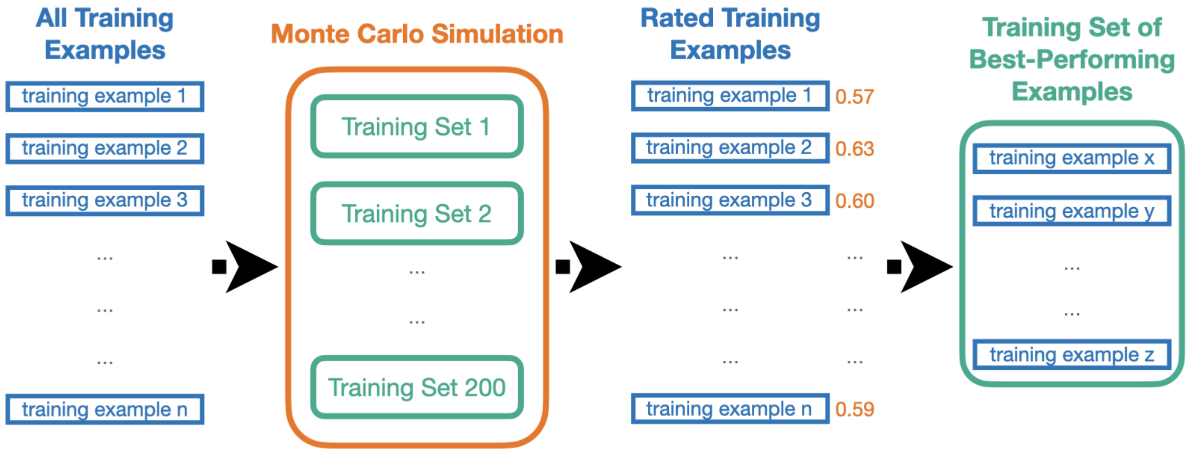

Figure 1: Overview of our proposed method. Starting from a small dataset of training examples, we perform a Monte Carlo simulation and calculate the average performance of each individual training example. Selecting the best-performing examples results in a training set more conducive to training.

whether a review is related or unrelated to the product. When a submitted review is unrelated to the product (e.g., a review that only mentions delivery time), the review is rejected.

Further, on our platform, multiple languages are covered across markets, with the majority of content written in German. Thus, **in our work, German is the focus language, with expansion to the in-domain lower-resource English and French**.

In synopsis, in an industry setting dealing with multiple markets and languages (e.g., an online shopping platform), (i) annotating large quantities of data for all languages is expensive and, (ii) the language of the dominant market makes up most of the available data. In line with these two observations, our contributions are:

1. showing that in few-shot settings, multilingual capabilities of large pretrained language models can be leveraged for better performance across languages,

2. proposing a Monte Carlo simulation method to identify training examples most conducive to learning based on a focus market (German), further improving overall performance.

We show that multilingual models finetuned on all languages perform better than their monolingual counterparts and that with our Monte Carlo selection method we can extract the training examples most conducive to learning to achieve improved performance, both in the monolingual and multilingual settings.

2 Related Work

It has been shown that large pretrained language models exhibit strong cross-lingual abilities, with cross-lingual transfer investigated extensively (Nooralahzadeh et al., 2020; K et al., 2020; Huang et al., 2019; Wu and Dredze, 2019; Pires et al., 2019; Conneau et al., 2020; Artetxe et al., 2020). In our work, we make use of cross-lingual transfer from higher- to lower-resource in-domain languages to improve performance.

With the emergent abilities of large language models (Wei et al., 2022), large models are being applied to few- and zero-shot settings (Sanh et al., 2022; Le Scao and Rush, 2021; Gao et al., 2021), showcased saliently in GPT-3, where prompting was shown to be effective across a range of tasks. To aid in few-shot learning, pattern-exploiting training (PET) was introduced in Schick and Schütze (2021a), allowing for training of large language models in few-shot settings via the use of prompts. It has been further shown that PET is competitive with models orders of magnitude larger (Schick and Schütze, 2021b).

Fu et al. (2022) showed that prompting can be employed in multilingual settings, with the authors showing that multilingual and multitask settings can be modeled without the use of language or task specific modules or training, by using prompts to leverage transfer learning capabilities. In our work we make use of prompts in a similar fashion, employing prompting to improve language transfer.

A downside with prompt-learning is the need for intricate and noisy prompt- and label-crafting, with

Language	Prompt
German	[MASK]: *r*
	r: Die Beurteilung ist [MASK]
	r ist [MASK]
English	[MASK]: *r*
	r: The review is [MASK]
	r is [MASK]
French	[MASK]: *r*
	r: L'avis est [MASK]
	r est [MASK]

Table 2: The prompts for each language, where *r* denotes the review text for each example

work on the task recently gaining more traction (Lu et al., 2022; Logan IV et al., 2022; Zhao and Schütze, 2021; Schick et al., 2020; Jung et al., 2022; Mishra et al., 2022; Wu et al., 2022; Shin et al., 2020). While there has been plenty of work in prompt- and label-crafting, there has been little work in the identification of optimal training sets. While training examples can have a large impact and add significant noise during training (due to the small size of the set), selecting examples most conducive to learning is still under-explored.

3 Experimental Setup

3.1 Pattern Exploiting Training (PET)

Patter Exploiting Training (Schick and Schütze, 2021a), or PET, is a technique that reformulates examples into cloze-type questions to help task finetuning of language models. It has been shown to be particularly effective in low-resource settings, outperforming setups with orders of magnitude more data (Schick and Schütze, 2021b).

In our work, we employ PET during the task finetuning phase, training a model to predict whether a review is *related* or *unrelated* to the product. Input examples are reformulated into prompts with exactly one masked token that the model learns to fill. The model output for that masked token is mapped to one of the task classes (in our case, *related* or *unrelated*). Formally, given model vocabulary V and classes C, the verbalizer maps $C \rightarrow V$. Input x is reformulated into input $x_p \in V^*$ with exactly one masked token.

Language	related	unrelated
German	unabhängig	verbunden
English	related	rejected
French	pertinent	mauvais

Table 3: The labels mapping to the two classes (i.e., the verbalizer) for each language

PET operates in three stages: (i) training a model for each prompt on a few annotated examples, (ii) soft-labeling a larger dataset of unlabeled data via an ensemble of prompt-trained models, (iii) training a final classifier on the soft-labeled dataset.

3.2 Models

Since PET works adjacently to Masked Language Modeling (MLM), models based on the `BERT`[1] architecture were chosen. For English, `bert-base-cased` was used. For German, we used the `bert-base-german-cased` variant. For French, we experimented with `FlauBERT` (Le et al., 2020) and `CamemBERT` (Martin et al., 2020). For `CamemBERT`, both the base and large variants were used.

As baselines we chose a logistic regression classifier and a production model for reviews moderation. The production model is based on the German `DistilBERT` model (Sanh et al., 2019). The model was originally finetuned on 20K reviews, 1271 of which were unrelated to the product and then further finetuned on 473 related and 198 unrelated reviews, to account for any shifts in data (e.g., temporal differences between original and testing data). This model is denoted with `DistilDE`.

In our work, since we have a plethora of reviews written in German, we focus on the German language for hyperparameter tuning and prompt labeling. Namely, we train using PET for 3 epochs, with a learning rate of 2e-3.

3.3 Prompt and Label Engineering

Prompts and labels were manually crafted for the German language, after empirically gauging their performance on the German development set, which is significantly larger than the English and French equivalents. We experimented both with the multi-token label variant (i.e., labels spanning multiple tokens) and multiple labels for a single

[1] Models as found on https://huggingface.co

class, both options were low-performing and we did not continue investigation in these directions.

Prompts for French and English were translated in a two-step process: (i) the word 'review' ('Beurteilung' in German, 'review' in English and 'avis' in French) was retrieved from the review module on the company website, which has been pre-translated by the localisation team, (ii) given the pre-translated word for 'review', the rest of the prompts were translated through Google Translate from German to the other two languages.

For labels, due to limitations with vanilla PET, we only chose words that span single tokens. For this reason, labels were hand-picked by the researchers to best approximate the essence of the class names, *related* and *unrelated*.

The prompts for each language are shown in Table 2. The verbalizer (label → class pairs) for each language is shown in Table 3.

3.3.1 Language Selection

In this work we focus on three languages: English, French and German. We do not verify whether customers are native speakers of these languages. Further, these languages cover multiple markets: French can be found in France, Luxembourg and Belgium, German in Austria, Germany and Switzerland, while English can be found in Germany, Ireland and the UK.

3.4 Data

Data comes from datasets of customer reviews on Zalando (an online fashion shopping platform). Reviews are submitted by customers and then moderated manually. Reviews are either accepted for publication or rejected because they do not meet the company's policy standards. When a review is rejected, it can be rejected for one or more reject reasons. These include offensiveness, divulging of personal data, and reviews unrelated to the product. Since most rejected reviews are reviews marked as unrelated to the product, we focus on this subclass since it promises the highest return of investment. Thus, the task we are solving is the binary classification between reviews related and unrelated to the product.

Zalando's largest market is German-speaking. Thus, (i) most reviews are written in German, (ii) there is an increased incentive to develop models to moderate German reviews. For these two reasons, we chose German as a 'focus' language for our work. Prompt and label engineering is conducted during experiments on the German set.

All data is made up of reviews submitted by customers after 2021 and up to June 2022, across products and product categories. For all languages, the training set (during prompt-based training) contains 8 related and 8 unrelated reviews. In English and French, the development set also contains 8 related and 8 unrelated reviews. In the German set, where more data is available, the development set is made up of 100 related and 100 unrelated reviews. For all languages, we collected 20,000 unlabeled reviews (to be soft-labeled during PET).

3.5 Monte Carlo Simulation

In few-shot settings, due to the natural scarcity of data, learning is particularly susceptible to noise in the training set. Performance relies heavily on the training set and minor, uninterpretable perturbations can affect performance drastically. Currently, selecting training examples is performed arbitrarily. We propose a method to identify which examples are most conducive to learning via a Monte Carlo simulation and selection of the examples that on average score the highest F1. The intuition behind this method is that a useful training example will on average be useful regardless of the other examples in the training set. Thus, with multiple runs, the useful training examples will on average score higher than the less useful examples.

Namely, we simulate 200 runs, via sampling 200 different training sets. For each set, we sample 16 reviews from a total of 32 possible reviews without repetition and without order-significance. The total number of combinations is intractably large. We instead sample 200 training sets due to computational considerations. Then, for each training set a model is trained using PET and evaluated on a common development set (and, finally, on the test set).

Due to the scarcity of data for English and French, the Monte Carlo simulation is performed solely on the German set. Performance was evaluated over a development set of 200 reviews (100 related and 100 unrelated to the product). The model we use is `bert-base-german-cased`.

4 Results

4.1 Monte Carlo Simulation

We performed 200 Monte Carlo runs, with `bert-base-german-cased` models trained using PET on

Method	Dev. F1	Accuracy	F1	Precision	Recall	tp, fp, fn, tn
Log. Reg.	-	72.2%	58.2%	97.5%	71.4%	1678, 44, 673, 184
DistilDE	-	88.8%	73.0%	96.8%	90.7%	2133, 70, 218, 158
Set$_{113}$	69.4%	**91.2%**	73.9%	95.5%	**94.8%**	**2228**, 104, **123**, 124
Set$_{031}$	70.3%	89.5%	71.8%	95.8%	92.5%	2175, 95, 176, 133
Set$_{054}$	71.0%	88.9%	70.2%	95.5%	92.2%	2168, 103, 183, 125
Set$_{148}$	72.7%	90.9%	74.1%	95.9%	94.1%	2212, 95, 139, 133
Set$_{111}$	74.9%	90.2%	73.4%	96.1%	93.1%	2188, 90, 163, 138
Set$_{full}$	48.6%	49.8%	43.6%	98.8%	45.3%	1068, *13*, 1283, *215*
Set$_{MC}$	-	89.8%	**74.7%**	97.1%	91.5%	2152, 65, 199, 163
mBERT$_{de}$	-	87.2%	60.1%	**92.9%**	93.0%	2186, 166, 165, 62
mBERT$_{all}$	-	82.3%	67.5%	**98.1%**	82.2%	1932, **37**, 419, **191**
mBERT$_{all/MC}$	-	86.0%	70.2%	97.2&	87.2%	2050, 59, 301, 169

Table 4: German performance comparison between sampled training sets performing the best on a development set versus the training set made up of the best-performing individual training examples. With bold we show the best score in each metric, except in the false positives and true negative columns, where the best performing model, Set$_{full}$, has not learned to recognize the positive class and thus has degenerate performance.

Method	Accuracy	F1	Precision	Recall	tp, fp, fn, tn
FlauBERT	60.6%	51.3%	88.1%	61.8%	1510, 204, 935, 243
CamemBERT$_{base}$	84.5%	45.8%	84.5%	100.0%	2445, 447, 0, 0
CamemBERT$_{large}$	83.9%	72.2%	92.6%	87.9%	2149, 171, 296, 276
mBERT$_{fr}$	**84.6%**	49.0%	84.9%	**99.3%**	**2431**, 432, **14**, 15
mBERT$_{all}$	77.1%	70.2%	**98.4%**	74.1%	1932, **29**, 633, **418**
mBERT$_{all/MC}$	82.5%	**74.4%**	96.5%	82.2%	2010, 72, 435, 375

Table 5: French test set performance.

sampled sets as detailed in Section 3.5. Models were evaluated on development and test sets with macro F1 score, Precision and Recall.[2] In Table 7 we see statistics on the performance of models. It is evident that performance is heavily reliant on the training set used in each iteration, with the difference between the minimum and maximum F1 scores being 40.5% in the development and 26.4% in the test set, with a standard deviation of 9.3% and 7.3% respectively. While mean performance for both sets is low (52.3% and 61.6% respectively), the maximum performance is high at 74.9% for the development and 74.1% for the test set.

We next investigate whether certain training examples are consistently more conducive to performance than other examples. In Table 8 we show the top- and bottom-3 ranked reviews based on their average F1 scores as calculated through the Monte Carlo simulation. While the worst-performing review contains multiple numbers (specifically, ‘36’ and ‘38’), which may inhibit learning, it is difficult to identify why the rest of the reviews perform better or worse. Further, reviews related and unrelated to the product are equally distributed as high- and low-performing. Nevertheless, a noticeable difference in performance can be observed, with a 4% absolute difference between the top- and bottom-scoring reviews. This exercise shows that while performance varies a lot across different training examples, it is difficult to infer why some examples perform better than others.

As a next step, we create a training set with the 16 reviews performing the best on the development set, by picking the 8 reviews related to the product with the highest score and the 8 unrelated reviews with the highest score. Combining these two sets of 8 reviews results in a 16-review training set to

[2] As implemented in https://scikit-learn.org.

Method	Accuracy	F1	Precision	Recall	tp, fp, fn, tn
BERT$_{en}$	83.2%	54.3%	98.5%	83.9%	4458, 67, 854, 101
mBERT$_{en}$	51.7%	36.1%	96.4%	52.1%	2769, 104, 2543, **64**
mBERT$_{all}$	89.7%	62.9%	**99.2%**	90.1%	4787, **39**, 525, 129
mBERT$_{all/MC}$	**93.6%**	**66.1%**	98.6%	**94.8%**	**5035**, 72, **277**, 96

Table 6: English test set performance.

Set	Min.	Max.	Mean.	Std
Dev.	34.4%	74.9%	52.3%	9.3%
Test	47.7%	74.1%	61.6%	7.3%

Table 7: Statistics of the Monte Carlo simulation performance on the development and test sets

be used in subsequent experiments. Models trained with this training set are marked with MC.

4.2 Unrelated Reviews Classification

We show that finetuning a single multilingual model on all available languages (English, French and German) outperforms its monolingual counterparts trained on each individual language. We can further improve performance by employing our proposed method of selecting training examples based on their average F1 score over a Monte Carlo simulation, improving performance in both the monolingual and multilingual settings.

For multilingual models, with mBERT$_x$ we denote the multilingual BERT model trained on the set of language x, for $x \in [fr, en, de]$. With mBERT$_{all}$ we denote the mBERT model finetuned on all languages, where the training set for German is the set that performed the best in the Monte Carlo experiments (Set$_{111}$). Finally, with mBERT$_{all/MC}$ we denote the variant of mBERT$_{all}$ where the German training set is made up of the training examples performing individually the best during the Monte Carlo experiments (the process is outlined in Section 3.5).

4.2.1 German Setting

With the greater availability of German data, this setting was chosen as the focus language of the project. While the other two languages (English and French), each have 16 reviews in the development set, there are 200 reviews available in German. For this reason, prompts and labels were crafted after evaluation on the German set (and subsequently translated into French and English) and the aforementioned Monte Carlo experiments (Section 4.1) for training example selection were performed on the German set. Results are shown in Table 4.

As baselines we compare against a logistic regression classifier trained on all 32 German reviews, as well as a production model based on German DistilBERT trained on 20K reviews.

As per the Monte Carlo experiments, 200 sets of 16 reviews were sampled from 32 total reviews, training bert-base-german-cased models on each set using PET. Here we show the five sets performing best on the development set, denoted with Set$_{xxx}$, where $x \in [0, 199]$. Set$_{MC}$ is created from the 16 best-performing training examples. For the multilingual transformer models, the training sets from English and French were used in conjunction with Set$_{111}$ (the best-performing set in the Monte Carlo experiments) forming mBERT$_{all}$, and with Set$_{MC}$ to form mBERT$_{all/MC}$.

To more fairly compare the multilingual models, which make use of more training examples (16 from each language, for 48 overall), we train a model with PET on all available German data (Set$_{full}$). This model does not seem able to generalize from all 32 examples, with very low performance when identifying reviews related to the product.

From the monolingual models, the best-performing one is Set$_{MC}$, trained on the training examples selected through the Monte Carlo simulation, outperforming the best-performing model trained during the Monte Carlo experiments (Set$_{111}$), with both a higher F1 score and lower false positive rate. Further, it outperforms the production model, DistilDE by almost 2%, despite requiring a fraction of training examples (16 versus 20K), with a slightly lower rate of false positives.

Between the multilingual models, the best-performing one is mBERT$_{all/MC}$, with an F1 score of 70.2% outperforming mBERT$_{all}$ and mBERT$_{de}$ with F1 scores 67.5% and 60.1% respectively. However, mBERT$_{all}$ has the lowest rate of false positives, with only 37 false positives versus 59 false

ID	Review	Avg. F1	Label
1	Sehr bequem und richtig süss 🥰 ich liebe baby rosa 😍🦋	63.9%	related
2	es ist ganz schade	63.9%	unrelated
3	Gutes Material auf der Haut, sitzt nicht ganz teilliert, etwas lockeres, hat es trotzdem behalten.	63.9%	related
4	Es würde ein anderes t-shirt schicken und zur nächsten würde ich wieder eine Retour machen.	60.0%	unrelated
5	Kann ich leider noch keine Bewertung abgegeben, ist noch bei Hermes wie gesagt	59.7%	unrelated
6	Ich musste eine Nummer größere bestellen, trage gewöhnlich 36. Ich hätte größe 38 kaufen sollen.	59.7%	related

Table 8: The top- and bottom-3 ranked reviews on average F1 performance from the Monte Carlo experiments for German (reviews here have been edited to preserve privacy and abide by GDPR laws).

positives from mBERT$_{all/MC}$. Nevertheless, the increase in the F1 score is significant at 2.7%, which shows that utilizing our proposed method of Monte Carlo selection and multilingual transfer performs the best for the multilingual setting too.

4.2.2 French Setting

For our experiments in French, we are comparing three monolingual models (FlauBERT, CamemBERT$_{base}$ and CamemBERT$_{large}$) with three multilingual models (i) mBERT$_{fr}$, (ii) mBERT$_{all}$, and (iii) mBERT$_{all/MC}$. Results are shown in Table 5.

Out of the three monolingual models, only CamemBERT$_{large}$ performs well, with FlauBERT having a low recall and CamemBERT$_{base}$ unable to identify reviews unrelated to the product. While CamemBERT$_{large}$ has a high F1 score, it suffers from a high rate of false positives, with 171 unrelated reviews classified as related. With our method this issue is mitigated, reducing the number of false positives to 29 and 72 with mBERT$_{all}$ and mBERT$_{all/MC}$ respectively, while at the same time keeping overall performance competitive or even better that the monolingual counterparts. While mBERT$_{all}$ performs slightly worse than CamemBERT$_{large}$, with an F1 score of 70.2% versus 72.2%, mBERT$_{all/MC}$ outperforms the monolingual model with an F1 score of 74.4%.

We can thus conclude that in French, performance improves both in reducing the false positive rate and in increasing the F1 score via multilingual transfer from the better-performing German training set selected through the Monte Carlo simulation (Set$_{MC}$) to the French set.

4.2.3 English Setting

For our experiments in English, we are comparing the monolingual BERT$_{en}$ model with three multilingual models (i) mBERT$_{en}$, (ii) mBERT$_{all}$, and (iii) mBERT$_{all/MC}$. Results are shown in Table 6.

While training mBERT solely on English does not perform well, with a very low recall score, mBERT$_{all}$ and especially mBERT$_{all/MC}$ perform well, showing that multilingual transfer and our proposed Monte Carlo selection method jointly improve performance. Namely, while the monolingual model has an F1 score of 54.3%, mBERT$_{all/MC}$ has an F1 score of 66.1%.

Unfortunately, in this case our proposed method does not provide improvements for the false positives rate. In fact, mBERT$_{all/MC}$ introduces 5 more false positives than the monolingual baseline. This could potentially be noise in the evaluation set, considering that English-language data was scarce, with only 168 unrelated reviews. Nevertheless, our proposed method improves the F1 score by 11.8% over the monolingual model.

4.2.4 Multilingual Setting

Finally, we compare in greater detail monolingual with multilingual model performance.

For our experiments in the multilingual setting, we compare three types of models. With BERT$_{mono}$ we denote the BERT model pretrained and finetuned on the corresponding language performing the best in each language setting (e.g., CamemBERT$_{large}$ for French), with mBERT$_{all}$ we denote the model trained on all language training sets and with mBERT$_{all/MC}$ we denote the multilingual model variant where the

Model	French	English	German
BERT_{mono}	72.2%	54.3%	**73.4%**
mBERT_{all}	70.2%	62.9%	67.5%
$\text{mBERT}_{all/MC}$	**74.4%**	**66.1%**	70.2%

Table 9: Comparison of monolingual and multilingual F1 scores per language on each test set.

Model	French	English	German
BERT_{mono}	38.3%	39.9%	39.5%
mBERT_{all}	**10.7%**	**23.2%**	**16.2%**
$\text{mBERT}_{all/MC}$	25.9%	42.9%	25.9%

Table 10: Comparison of monolingual and multilingual false positive rate per language on each test set.

German Set_{MC} was used instead of Set_{111}.

The monolingual models are the models that performed the best in their respective language settings on the development sets (although for French and English the development sets contain only 16 reviews). The multilingual models were all finetuned using English labels and prompts, since English data is the most prominent in mBERT's pretraining set and intuitively the best hub across languages (even though Anastasopoulos and Neubig (2020) showed that choosing English is not always the best hub-language for bilingual models, hub-language selection is out of scope for our work).

In Table 9 we compare F1 scores. In French and English, multilingual models perform best, with an average increase to the F1 score of 2% for mBERT_{all} and 11% for $\text{mBERT}_{all/MC}$. In German, while $\text{mBERT}_{all/MC}$ performs competitively, the monolingual model still performs the best. This is to be expected, considering that German was the focus language of our experiments and the German monolingual models received the majority of attention during the development stage, with hyperparameter tuning and extensive prompt engineering. On the other hand, no tuning or engineering was performed on the multilingual models.

In Table 10, we compare false positive rates (i.e., unrelated reviews that were classified as related) between monolingual and multilingual models. In many real-world settings, false positives are particularly insidious, since users are exposed to content they should not be seeing. For example, exposing customers to harmful content or publishing reviews where customers have inadvertently revealed personal information (e.g., their address or email address) is harmful. In our case, customers are exposed to information that is not related to the product, which may add confusion and affect customer trust negatively.

For this important metric, we see that the monolingual models perform badly, with the false positive rate just below 40% across all languages. This is performance that would deem such models untrustworthy for production. On the other hand, the multilingual models perform better overall. In particular, mBERT_{all} performs the best in all languages and by large margins (at least 16%). $\text{mBERT}_{all/MC}$ also performs better than the monolingual models in French and German, sporting improvements of at least 12%. In English, however, $\text{mBERT}_{all/MC}$ performs competitively but still worse than the monolingual model by 3%. An explanation for this is that in English the test set contains only a few unrelated (i.e., negative) reviews, numbering at 168. In comparison, in the German set there are 228 unrelated reviews and in the French set 447. Thus, due to the small size, it is challenging to make inference on solely unrelated reviews.

5 Conclusion

In our work, we investigate how to improve discrimination between customer reviews related and unrelated to the product in low-resource settings for English, French and German. We show that via multilingual transfer learning we can improve performance of models in English and French, leveraging the in-domain higher-resource German data, while at the same time reducing the rate of false positives across all languages.

Selecting training examples most conducive to performance in few-shot learning is of paramount importance and still an open question. We propose a method to extract such examples through a Monte Carlo simulation, selecting training examples with the highest average performance across experiments. We show that with our method we can improve performance both in monolingual and multilingual settings, outperforming baselines with orders of magnitude more data as well as all models trained on randomly-sampled sets, consistently increasing F1 scores and decreasing false positives.

Acknowledgements. The first author performed work as an intern at Zalando SE (Berlin). The first author is currently funded by the European Research Council (grant #740516).

6 Limitations

In our Monte Carlo simulation, due to computational restrictions, we only performed 200 random runs. Even though the training set produced via our selection method outperforms other methods, performance could be potentially improved even further with more runs.

Future work should focus on expanding the set of languages investigated. Due to limited data resources, we were not able to procure enough data in other languages. It is important to include not only more, but also more linguistically diverse languages in the study.

References

Antonios Anastasopoulos and Graham Neubig. 2020. Should all cross-lingual embeddings speak English? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8658–8679, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot prompt: Multilingual multitask promptraining.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods* *in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.